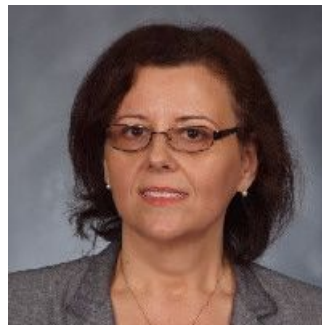


Dec 2018 DataOne Webinar

A Shared Staffing Model for Research Data Curation



Lisa Johnston
University of Minnesota



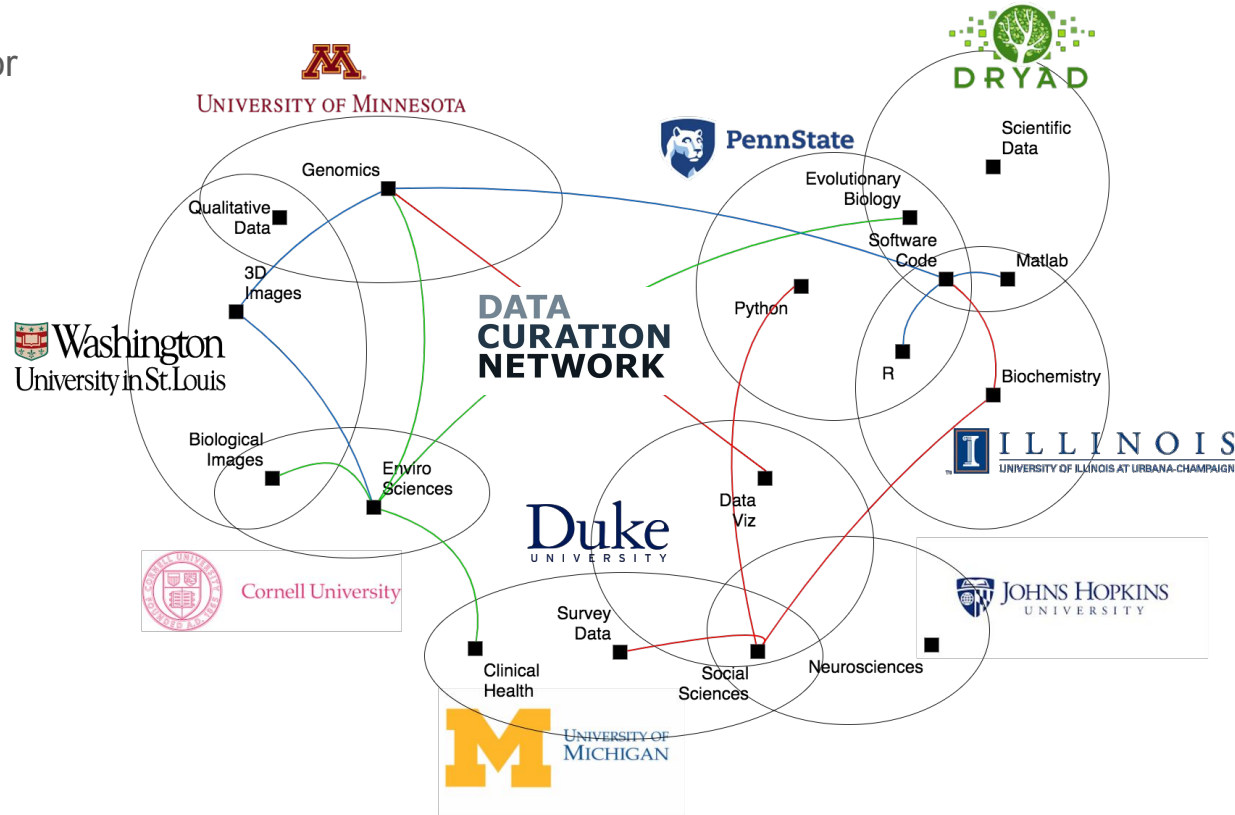
Debra Fagan
Dryad Data Repository

2018-12-11

DataCurationNetwork.org

DATA CURATION NETWORK

- Collaborative staffing model for curating research data across academic and general data repositories
- Funded by Alfred P. Sloan Foundation
- Implementation phase (2018-2021)
 - 8 member institutions
 - 32 individuals
- Goal is to expand to all users in 2020



Well curated data are more valuable.

- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted and reused
- The research findings they represent are more likely to be validated and replicated
- More likely to be properly cited

Due to the **heterogeneous and multidisciplinary** nature of research data generated by the researchers we support, data curation can be a challenge.

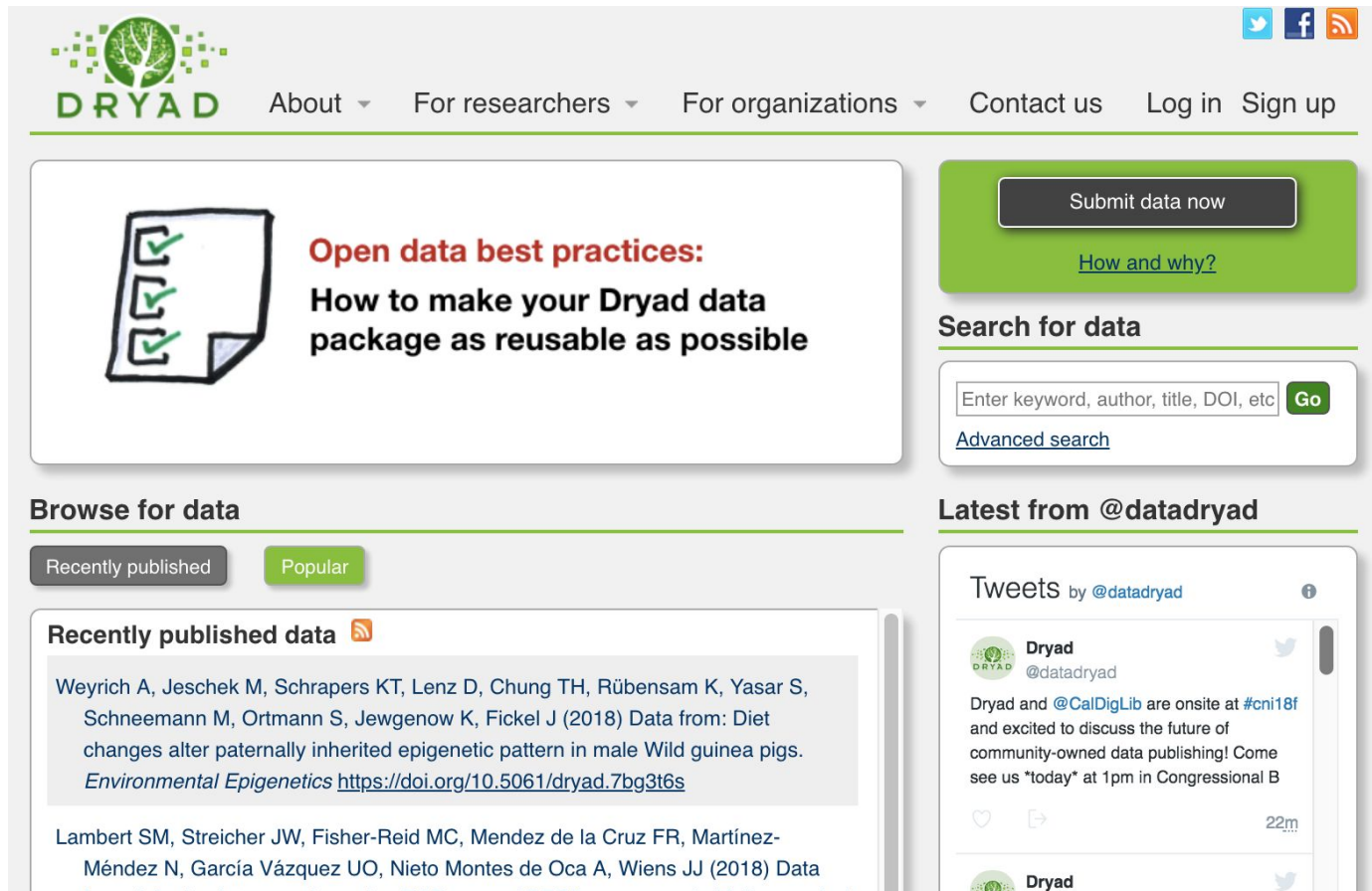
The Data Curation Network (DCN)
addresses this challenge by
collaboratively sharing data curation staff
across a network of partner institutions and data
repositories.

DRYAD

Nonprofit, curated
general-purpose
repository for data
underlying scholarly
publications

Goal - make data
discoverable, freely
reusable and citable

Evolved from ecology
and evolutionary bio
data -> diverse data in
a variety of disciplines



The screenshot shows the DRYAD website homepage. At the top is the DRYAD logo, a green tree-like icon with the word "DRYAD" below it. To the right of the logo are navigation links: "About", "For researchers", "For organizations", "Contact us", "Log in", and "Sign up". In the top right corner are social media icons for Twitter, Facebook, and RSS. Below the navigation bar is a large white box with a green border. On the left is an icon of a document with three green checkmarks. To the right of the icon is the text: "Open data best practices: How to make your Dryad data package as reusable as possible". To the right of this box is a green button that says "Submit data now" and a link "How and why?". Below this is a search bar with the placeholder text "Enter keyword, author, title, DOI, etc" and a green "Go" button. Below the search bar is a link "Advanced search". Below the search bar is a section titled "Browse for data" with two buttons: "Recently published" and "Popular". Below the "Recently published" button is a section titled "Recently published data" with a red RSS icon. Below this section are two entries of recently published data. The first entry is by Weyrich A, Jeschek M, Schrapers KT, Lenz D, Chung TH, Rübensam K, Yasar S, Schneemann M, Ortmann S, Jewgenow K, Fickel J (2018) Data from: Diet changes alter paternally inherited epigenetic pattern in male Wild guinea pigs. *Environmental Epigenetics* <https://doi.org/10.5061/dryad.7bg3t6s>. The second entry is by Lambert SM, Streicher JW, Fisher-Reid MC, Mendez de la Cruz FR, Martínez-Méndez N, García Vázquez UO, Nieto Montes de Oca A, Wiens JJ (2018) Data. Below the "Browse for data" section is a section titled "Latest from @datadryad" with a Twitter icon. Below this section is a tweet from Dryad (@datadryad) that says: "Dryad and @CalDigLib are onsite at #cni18f and excited to discuss the future of community-owned data publishing! Come see us *today* at 1pm in Congressional B". The tweet has 22 replies.

DRYAD

About For researchers For organizations Contact us Log in Sign up

Submit data now

[How and why?](#)

Search for data

Enter keyword, author, title, DOI, etc Go

[Advanced search](#)

Browse for data

Recently published Popular

Recently published data

Weyrich A, Jeschek M, Schrapers KT, Lenz D, Chung TH, Rübensam K, Yasar S, Schneemann M, Ortmann S, Jewgenow K, Fickel J (2018) Data from: Diet changes alter paternally inherited epigenetic pattern in male Wild guinea pigs. *Environmental Epigenetics* <https://doi.org/10.5061/dryad.7bg3t6s>

Lambert SM, Streicher JW, Fisher-Reid MC, Mendez de la Cruz FR, Martínez-Méndez N, García Vázquez UO, Nieto Montes de Oca A, Wiens JJ (2018) Data

Latest from @datadryad

Tweets by @datadryad

Dryad @datadryad

Dryad and @CalDigLib are onsite at #cni18f and excited to discuss the future of community-owned data publishing! Come see us *today* at 1pm in Congressional B

22m

Dryad

Dryad - Curation Challenges

Numerous challenges:

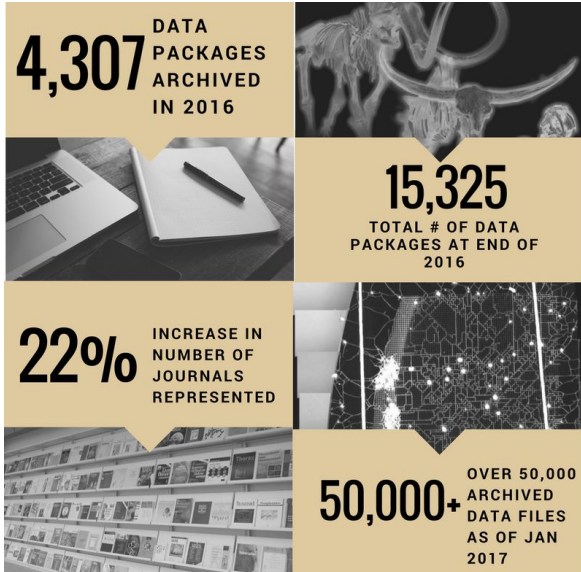
Licensing / Copyright

Human subject information

Endangered species

Discipline specific knowledge

Skills and resources needed to curate a wide variety of data



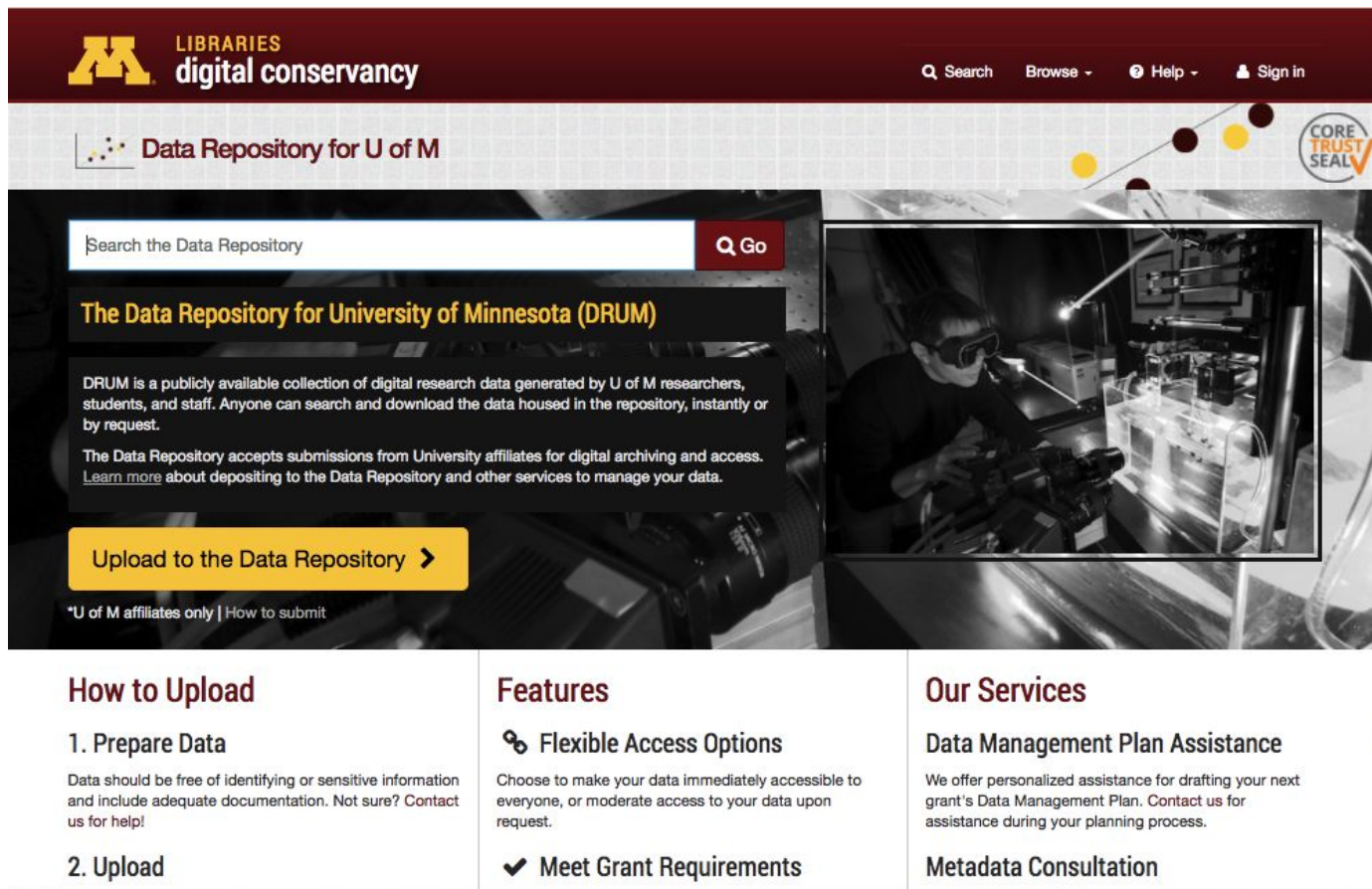
DRUM

Mediated deposit
with post-ingest
curation

7 data curation staff

- Scientific (2)
- Spatial/GIS
- Public Health
- Digital Hum
- Social Sciences
- Coordinator

Part of a larger RDS
service run by the
libraries



LIBRARIES digital conservancy

Search Browse Help Sign in

Data Repository for U of M

Search the Data Repository **Go**

The Data Repository for University of Minnesota (DRUM)

DRUM is a publicly available collection of digital research data generated by U of M researchers, students, and staff. Anyone can search and download the data housed in the repository, instantly or by request.

The Data Repository accepts submissions from University affiliates for digital archiving and access. [Learn more](#) about depositing to the Data Repository and other services to manage your data.

Upload to the Data Repository

*U of M affiliates only | [How to submit](#)

How to Upload

- 1. Prepare Data**
Data should be free of identifying or sensitive information and include adequate documentation. Not sure? Contact us for help!
- 2. Upload**

Features

- Flexible Access Options**
Choose to make your data immediately accessible to everyone, or moderate access to your data upon request.
- Meet Grant Requirements**

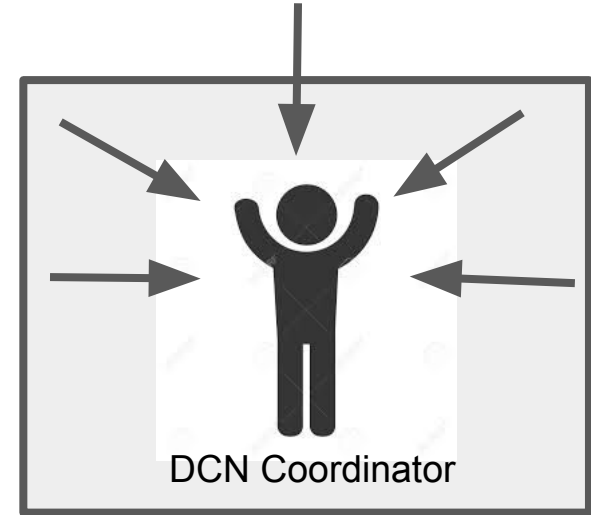
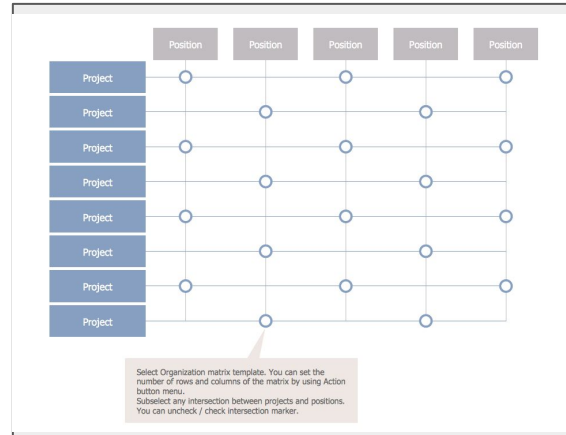
Our Services

- Data Management Plan Assistance**
We offer personalized assistance for drafting your next grant's Data Management Plan. Contact us for assistance during your planning process.
- Metadata Consultation**

Project → Service

1. How do we get the individual datasets to the right data curation expert?
2. How do we account for differences across institutions with differing policies/infrastructure?
3. How do we create consistency across different curators?
4. How do we communicate effectively?
5. How do we sustain and grow?

How do we get the individual datasets to the right data curation expert?



How do we account for differences across institutions with differing policies/infrastructure?



Understand your partners! Baseline assessment and researcher focus groups at each institution

Journal of eScience Librarianship

ISSN 2161-3974

Volume 6 | Issue 1

Article 3

2017-02-28

Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services



**Journal of Librarianship and
Scholarly Communication**

Editing: How Important is Data Curation? Gaps and Opportunities for Academic Libraries

Share: [f](#) [t](#) [g+](#) [in](#)

Research Article

How Important is Data Curation? Gaps and Opportunities for Academic Libraries

Authors: Lisa R Johnston ✉, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart

Unique Perspectives and Infrastructure



Institution	Data Repository	Launch Date	Dataset Holdings (as of April 2018)	Technology Platform
Cornell University	eCommons	Fall 2002	120 datasets	DSpace 6.2
Dryad Digital Repository	Dryad Digital Repository	2009	21,293 data packages; 67,907 individual files	DSpace moving to CDL's Merritt
Duke University	Duke Digital Repository (DDR)	Jan 2017	34 data deposits total: 8,120 files (most average 50-100)	Fedora 3.8, and Hyrax
Johns Hopkins University	JHU Data Archive	2011	61 datasets or dataverses; 448 files	Dataverse
Penn State University	ScholarSphere	Fall 2012	966 files	Hydra/Fedora (soon to be Sufia 7)
University of Illinois	Illinois Data Bank	May 2016	77 datasets	Ruby on Rails interface with our preservation repository (Medusa)
University of Michigan	Deep Blue Data	Sep 2016	114 datasets	Samvera/Fedora Hyrax
University of Minnesota	Data Repository for the U of M (DRUM)	Mar 2015	199 in DRUM, ~600 in IR	DSpace 5.5

How do we account for differences across institutions with differing policies/infrastructure?

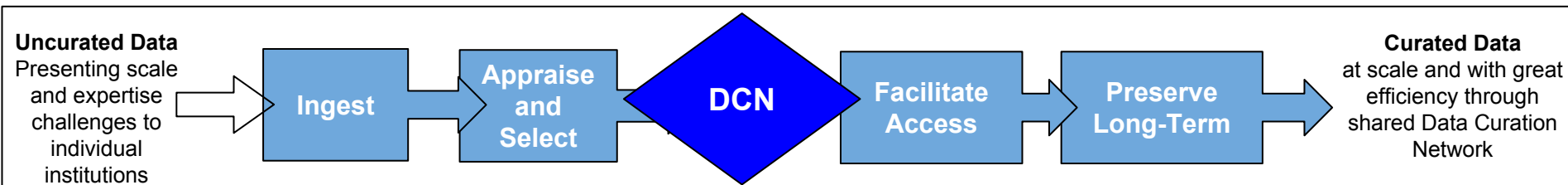
1

Understand your partners! Baseline assessment and researcher focus groups at each institution

2

DCN workflow with built-in flexibility for local interpretation

DCN Workflow

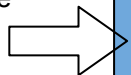


- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)

DCN Workflow

Uncurated Data

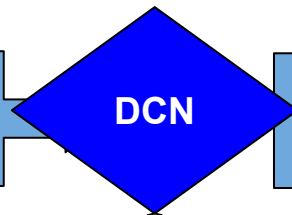
Presenting scale and expertise challenges to individual institutions



Ingest



Appraise and Select



DCN

Facilitate Access



Preserve Long-Term

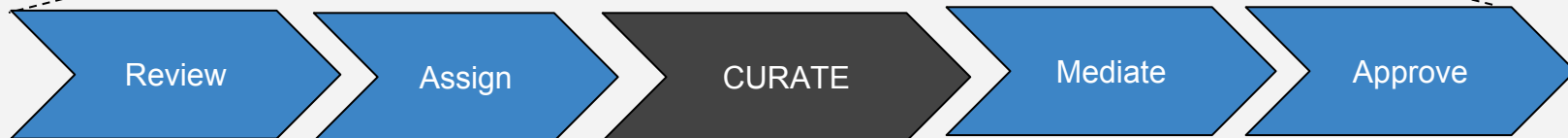


Curated Data

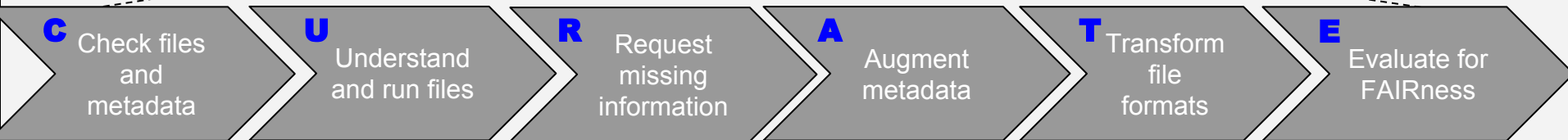
at scale and with great efficiency through shared Data Curation Network

Data Curation Network

DCN Coordinator Workflow

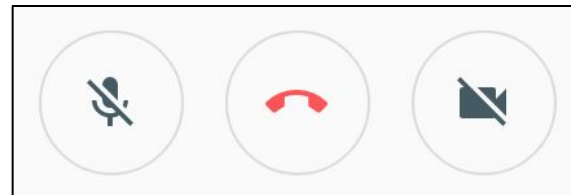
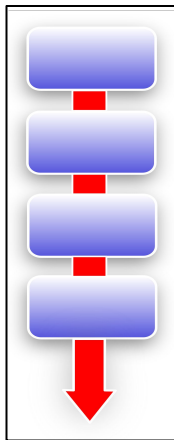


DCN Curator Workflow



How do we account for differences across individual curators?

CURATE



CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- C** **Check** data files and read documentation
- U** **Understand** the data (try to), if not...
- R** **Request** missing information or changes
- A** **Augment** the submission with metadata for findability
- T** **Transform** file formats for reuse and long-term preservation
- E** **Evaluate** and rate the overall submission for FAIRness.

DCN CURATE Checklists

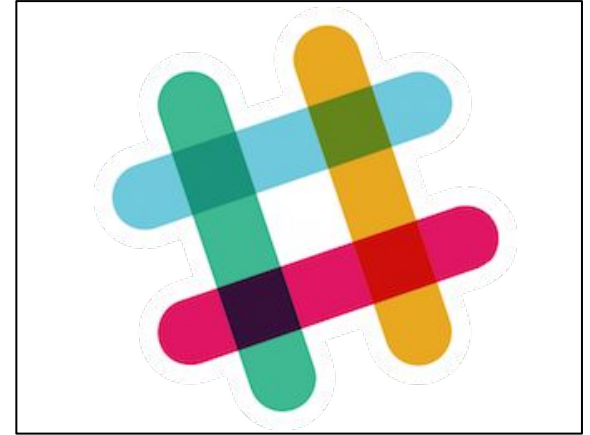
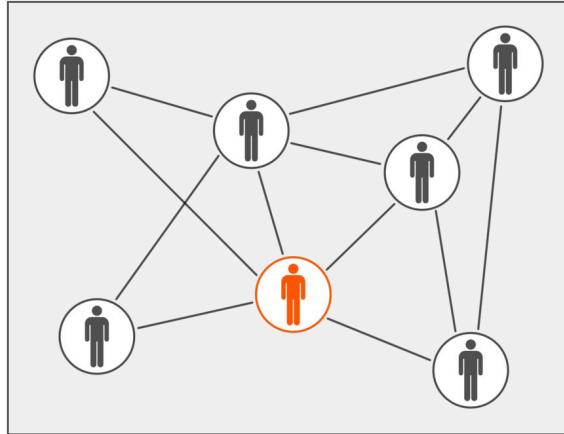
Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

CURATE Actions	Curation Check	Evaluate and rate the overall data record for FAIRness. ²	Findable -
Check data files and read documentation <ul style="list-style-type: none"> Review the content of the data files (e.g., open and run the files or code). Verify all metadata provided by the author and review the available documentation. 	<input type="checkbox"/> Files open as expected <input type="checkbox"/> Issues _____ <input type="checkbox"/> Code runs as expected <input type="checkbox"/> Produces minor errors <input type="checkbox"/> Does not run and many errors <input type="checkbox"/> Metadata quality is rich, complete <input type="checkbox"/> Metadata has issues <input type="checkbox"/> Documentation Type (citation, Readme / Codebook / Data dictionary / Other: _____) <input type="checkbox"/> Missing/None <input type="checkbox"/> Needs work	<ul style="list-style-type: none"> Score the dataset and recommend ways to increase the FAIRness of the data and become “DCN approved.” 	<input type="checkbox"/> Metadata exceeds author/ title/ date, Unique PID (DOI, Handle, PURL, etc.). <input type="checkbox"/> Discoverable via web search engines like Google.
			Accessible - <input type="checkbox"/> Retrievable via a standard protocol (e.g., HTTP). <input type="checkbox"/> Free, open (e.g., download link).
			Interoperable - <input type="checkbox"/> Metadata formatted in a standard schema (e.g., Dublin Core). <input type="checkbox"/> Metadata provided in machine-readable format (OAI feed).
			Reusable - <input type="checkbox"/> Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author. <input type="checkbox"/> Clear indicators of who created, owns, and stewards the data. <input type="checkbox"/> Data are released with clear data usage terms (e.g., a CC License).
Understand the data (or try to) <ul style="list-style-type: none"> Check for quality assurance and usability issues such as missing 	<i>Varies based on file formats and example....</i>		

¹ Format Recommendations, <http://guides.library.cornell.edu/eccommons/formats>

² Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

How do we communicate effectively?



How do we sustain and grow?

1

Train new curators (IMLS grant!)

DATA CURATION WORKSHOP

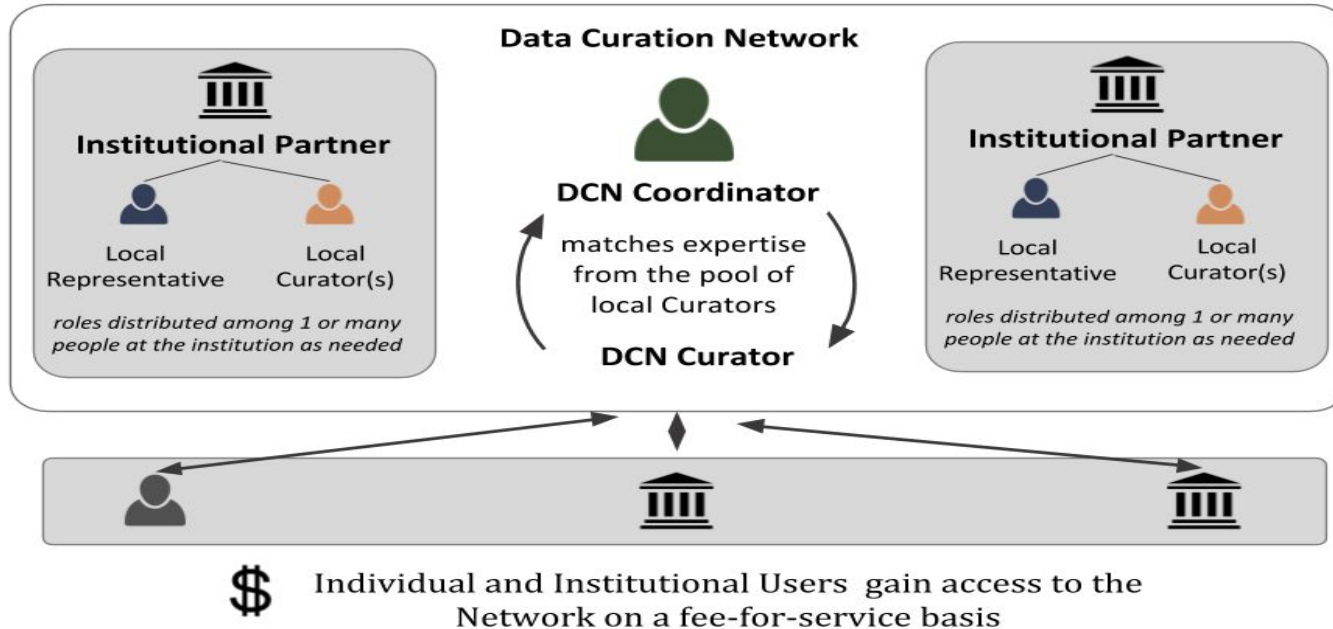


How do we sustain and grow?

Alliance Model

Institutional Partners contribute staff and fund central coordinator

2



DCN Research Questions

Is a networked approach to curating research data more efficient?

- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)
- Curator incentives?

Are curated data are more valuable?

- Survey researcher satisfaction
- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.

Thanks!

<https://DataCurationNetwork.org>

Twitter: #DataCurationNetwork